上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

# 数据挖掘与商务分析

# 第2讲 数据预处理

主讲教师：肖升生

# 课程导入

- 从自身的研究经历说起……

boardeX
exterior sheathing

VS.

# 讲授提纲

01 数据预处理概述

02 数据清理

03 数据抽样

04 数据规约

05 数据变换

06 数据预处理案例

# 讲授提纲

01 数据预处理概述

02 数据清理

03 数据抽样

04 数据规约

05 数据变换

06 数据预处理案例

# 数据集形式

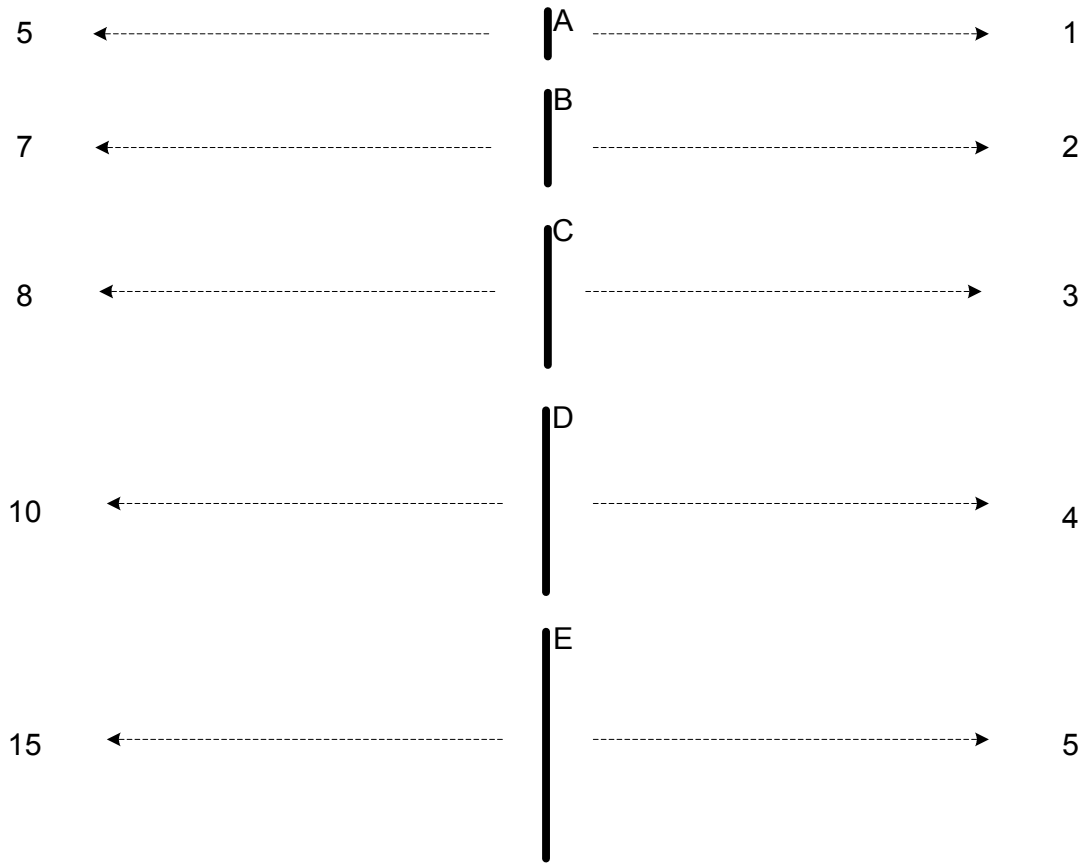
列：属性、特征、变量

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

行：观测、样本点、记录

# 属性的测度和取值

- 属性和属性取值之间是有区别的
- 属性的取值与属性测度



# 属性类型

- 标称属性（Nominal）
  - 如身份证号码, 邮编，国籍
- 序数属性（Ordinal）
  - 如成绩等级{A,B,C}、收入水平{高、中、低}, 教师职称{教授、副教授、助理教授}
- 区间属性（Interval）
  - 如日期、摄氏温度
- 比率属性（Ratio）
  - 如长度、价格、数目

# 属性类型判定

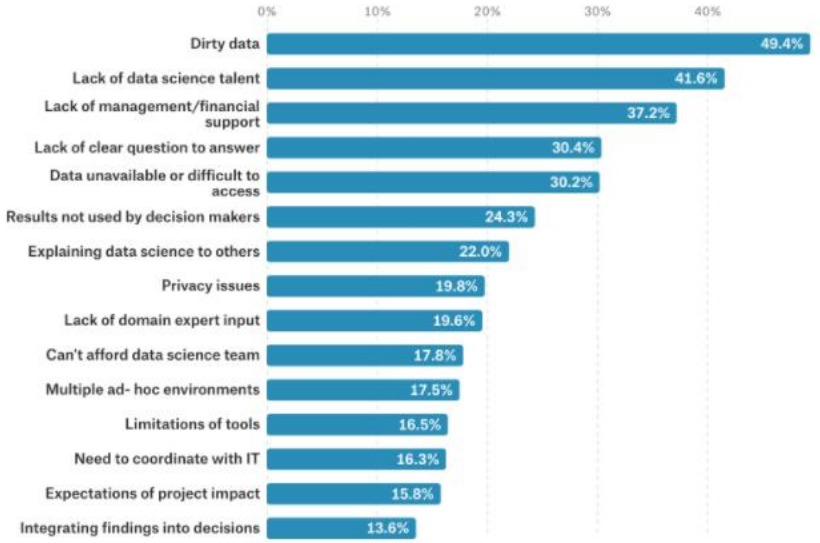
- 属性类型的判定取决于能否进行下列运算：
  - 取等（Distinctness）：　＝　≠
  - 排序（Order）：　＜　＞
  - 加减（Addition）：　＋　－
  - 乘除（Multiplication）：　×　÷

  - 标称属性：取等
  - 序数属性：取等、排序
  - 区间属性：取等、排序、加法
  - 比率属性：取等、排序、加法、乘法

# “脏”数据

- 现实世界的数据是“脏”的
  - 不完整
  - 含噪声
  - 不一致

Kaggle调查：机器学习人员在工作中的最大障碍?




- 数据为什么会变“脏”？
  - 收集缺乏合适值的数据、人为/硬件/软件问题……
  - 数据收集工具、数据输入时人为错误、传输中产生的错误……
  - 时空不一致的数据源、冗余数据的修改不完全……

https://www.vidora.com/general/kaggle-survey/

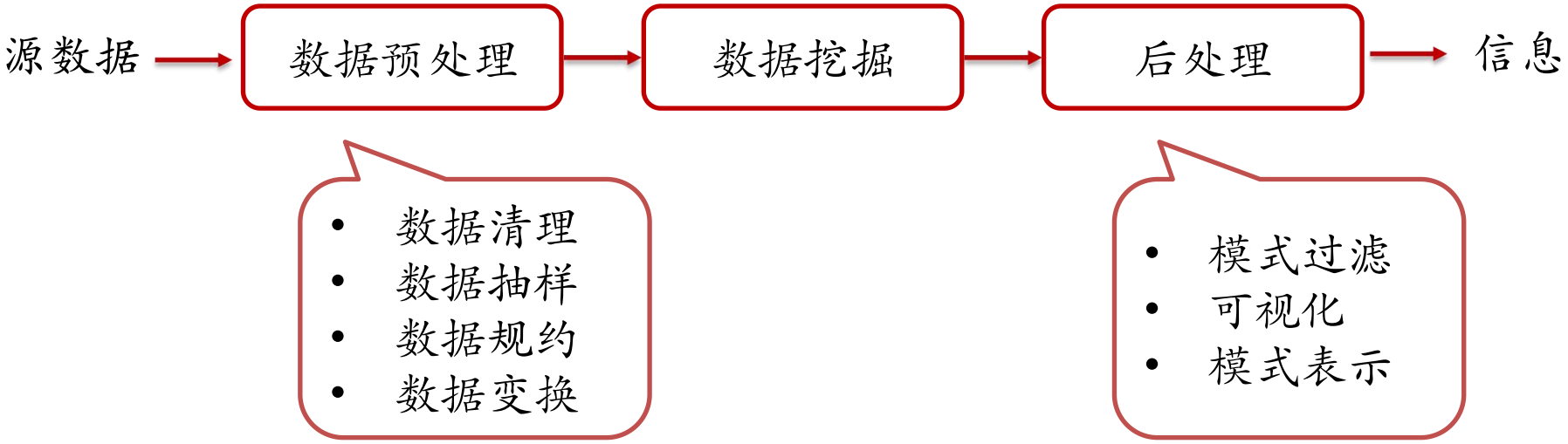
# 数据质量的多维度量

- **准确性**（Accuracy）：数据的记录状态能够准确表示物理世界实际状况的程度；
- **完整性**（Integrity）：描述特征、特征属性和特征关系的完善程度；
- **一致性**（Consistency）：数据集记录、格式、内容等方面的一致程度；
- **时效性**（Timeliness）：描述数据的更新频次对数据质量的影响；
- 可信性（Believability）：描述数据的正确性的可信程度；
- 可解释性（Interpretability）：描述数据被理解的程度；

……

# 数据预处理的作用

- 原因：现实世界的大部分数据质量不高，存在不完整、不一致、不准确等问题；
- 作用：数据预处理可以改进数据质量，有助于提高数据挖掘的准确率和效率。



# 数据预处理的主要步骤

- 数据清理（Data Cleaning）
  - 把“脏”数据变成“干净的”，提升数据质量；
  - 如处理缺失值，光滑噪声数据，处理离群点，解决不一致性等；
- 数据抽样（Data Sampling）
  - 抽取有代表性的数据记录，进行挖掘分析；

行的选取

- 数据规约（Data Reduction）
  - 寻找针对目标任务的有用特征，以缩减数据规模，并达到基本一致的挖掘效果；
  - 维规约、属性子集选择；

列的精简

- 数据变换（Data Transformation）
  - 通过特定函数将原属性取值映射到新的函数空间。
  - 规范化、离散化、简单函数变换

值的映射

# 讲授提纲

# 数据清理

- 目标：把“脏”数据变成“干净的”，提升数据质量；

- 数据清理的主要任务
  - **处理缺失值；**
  - **光滑噪声数据；**
  - **识别或删除离群点；**
  - 解决数据中的不一致性；
  - 删除重复数据；

# 缺失值的类型

- 完全随机缺失（Missing completely at random）
  - 某字段是否缺失与自身和其他观测变量无关，是完全随机的；
  - 比如机器完全随机采样造成；
- 随机缺失（Missing at random）
  - 某字段缺失的概率不是完全随机的，可以由其他观测变量解释；
  - 比如问卷调查中，收入数据的缺失可能与受访者性别有关；
- 不随机缺失（Missing not at random）
  - 字段缺失的概率与该字段本身的取值有关；
  - 比如收入数据的缺失可能与受访者本身收入高低有关；

# 处理缺失值

- 删除整条观测数据
  - 当类别数据缺失时通常这么做；
  - 适合数据集中缺失值较少的情况；
- 人工填补缺失值：工作量大，可行性低
- 自动填补缺失值
  - 使用一个全局变量：比如unknown或0
  - 使用属性的中心度量：平均值、中位数
  - 使用最可能的值填充：基于贝叶斯模型或决策树自动推断

# 噪声

- 噪声是测量误差的随机部分，如值被扭曲或加入了偏误；

(a) 时间序列

(b) 加入噪声的时间序列

思考：如果噪声偏误进一步增大呢？

# 光滑噪声数据

- 分箱
  - 首先排序数据，并将他们分到等深的箱中；
  - 然后按箱的平均值、中值或边界值平滑数据。
- 回归
  - 通过使数据拟合一个回归函数来平滑数据；
  - 试图发现相关变量之间的变化模式；
- 聚类
  - 通过聚类来检测并删除离群点；

**通过分箱平滑数据**

划分为（等深的）箱：

箱1：4，8，15

箱2：21，21，24

箱3：25，28，34

用箱平均值平滑：

箱1：9，9，9

箱2：22，22，22

箱3：29，29，29

用箱边界平滑：

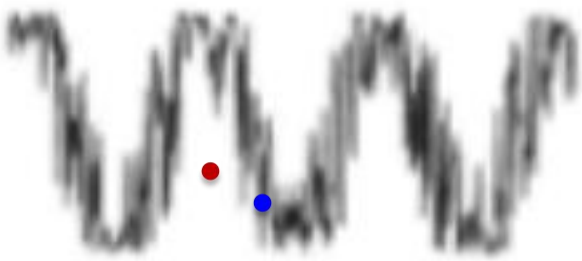
箱1：4，4，15

箱2：21，21，24

箱3：25，25，34

# 离群点

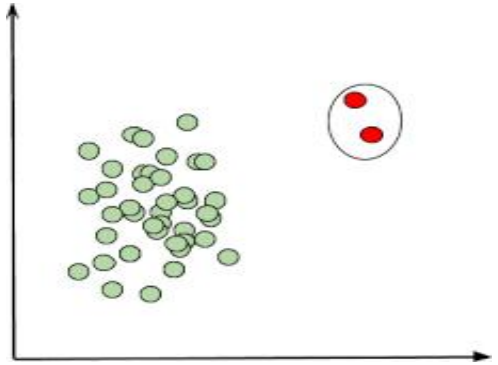
- 离群点（outliers）：不同于数据集中大部分对象取值的特殊对象，也称为异常对象或异常值；
- 区分噪声和离群点
  - 噪声：不正确的、偏误数据；
  - 离群点：取值上“异于常人”；
  - 关系：离群点可能是噪声，也可能是真实、正确的数据；



- 离群点检测的应用
  - 信用风险监测、网络入侵检测等

# 离群点检测方法

- 基于模型的技术
  - 建立数据表征模型，模型不能完美拟合的对象为“离群点”；
- 基于邻近度的技术
  - 定义对象之间的临近性度量，远离大部分其他对象的则是“离群点”；
- 基于密度的技术
  - 通过数据估计对象分布的密度，分布密度明显低于其他部分的称之为“离群点”。



详见《数据挖掘导论》第10章异常检测

# 小结

- 数据清理的目标
  - 把“脏”数据变成“干净的”，提升数据质量；
- 数据清理的主要任务
  - **处理缺失值**:删除、统一插补、自动推断；
  - **光滑噪声数据**：分箱、回归、聚类；
  - **识别或删除离群点**：基于模型、邻近度、密度的技术
  - 解决数据中的不一致性；
  - 删除重复数据；

# 讲授提纲

01 数据预处理概述

02 数据清理

03 **数据抽样**

04 数据规约

05 数据变换

06 数据预处理案例

# 数据抽样导入

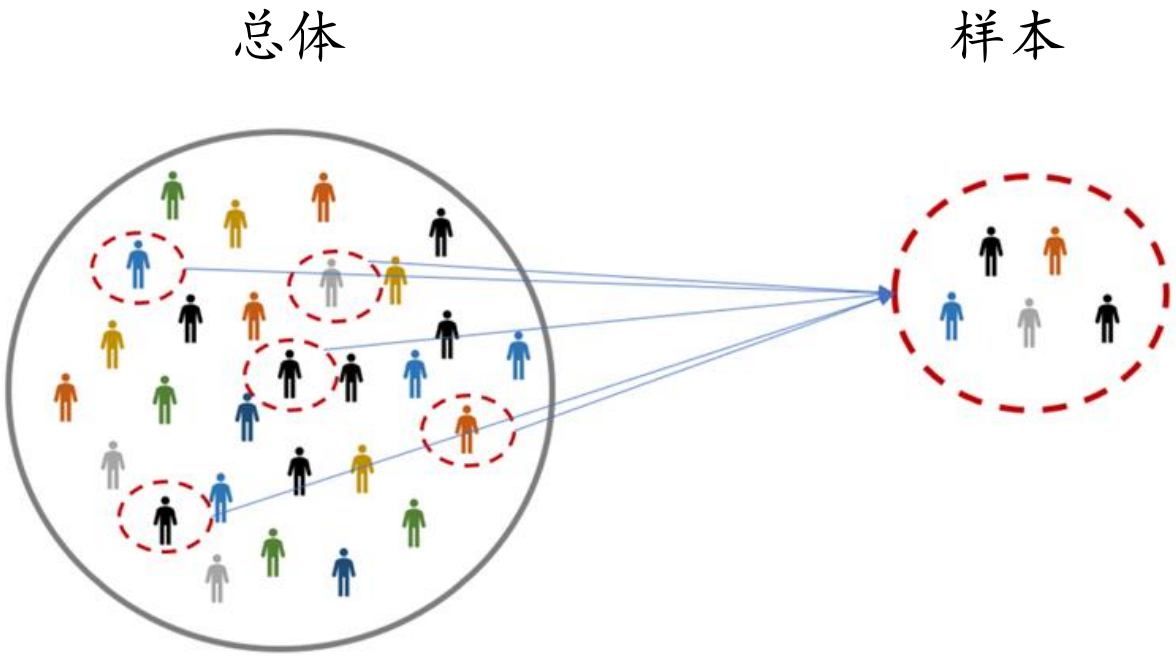
一天爸爸叫儿子去买一盒火柴，临出门前爸爸嘱咐儿子要买能划燃的火柴。儿子拿着钱出门了。过了一会儿，儿子才回到家。

“火柴能划燃吗？” 爸爸问。

“都能划燃。” 儿子递过一盒划过的火柴，兴奋地说：“我每根都试过了。”

# 抽样的思想

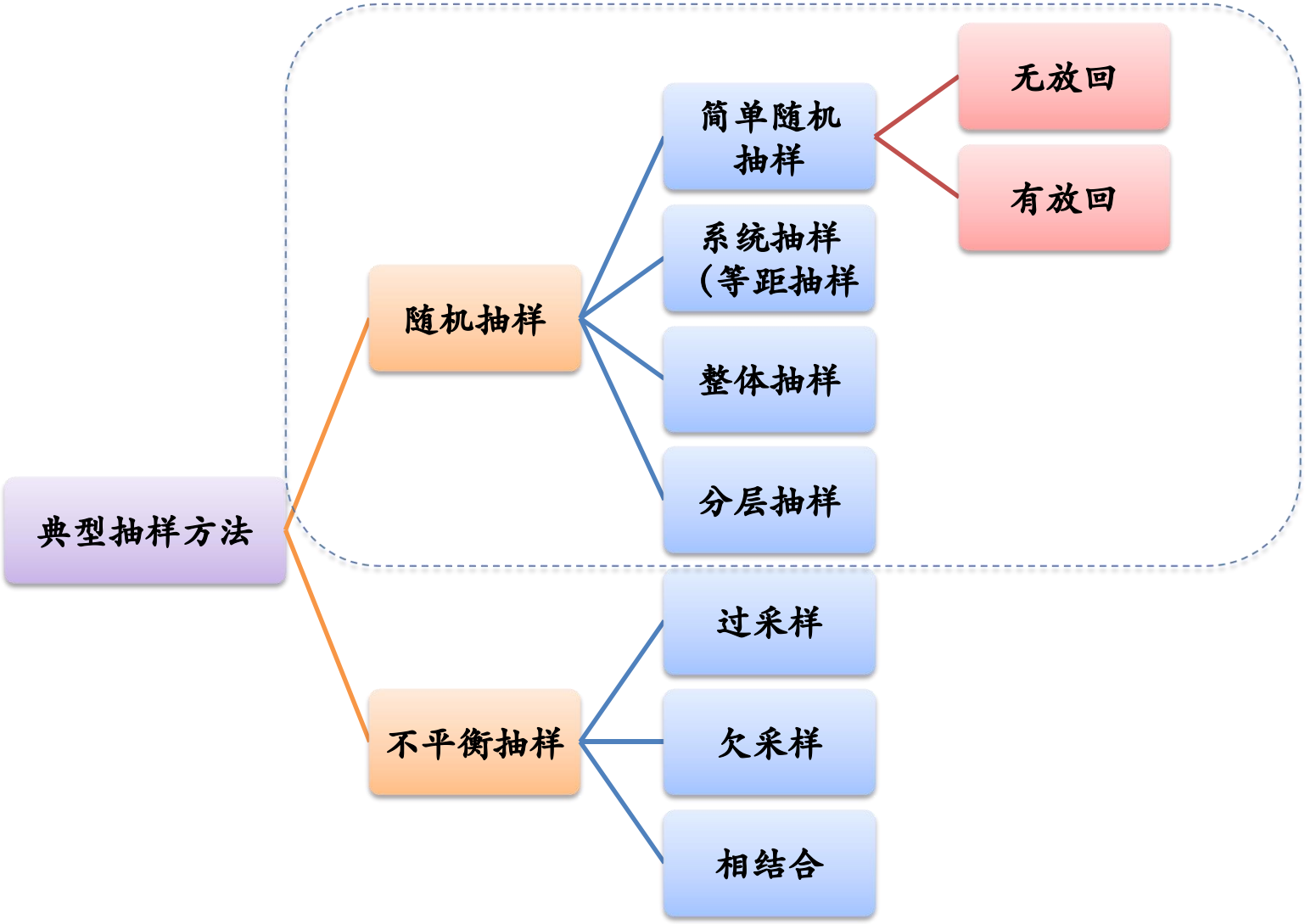
- 用样本估计总体
  - 即通常不直接去研究总体，而是通过从总体中抽取一部分样本，根据样本的情况去估计总体的相应情况。
- 关键原则：选取有**代表性**的子集；



# 数据抽样的作用

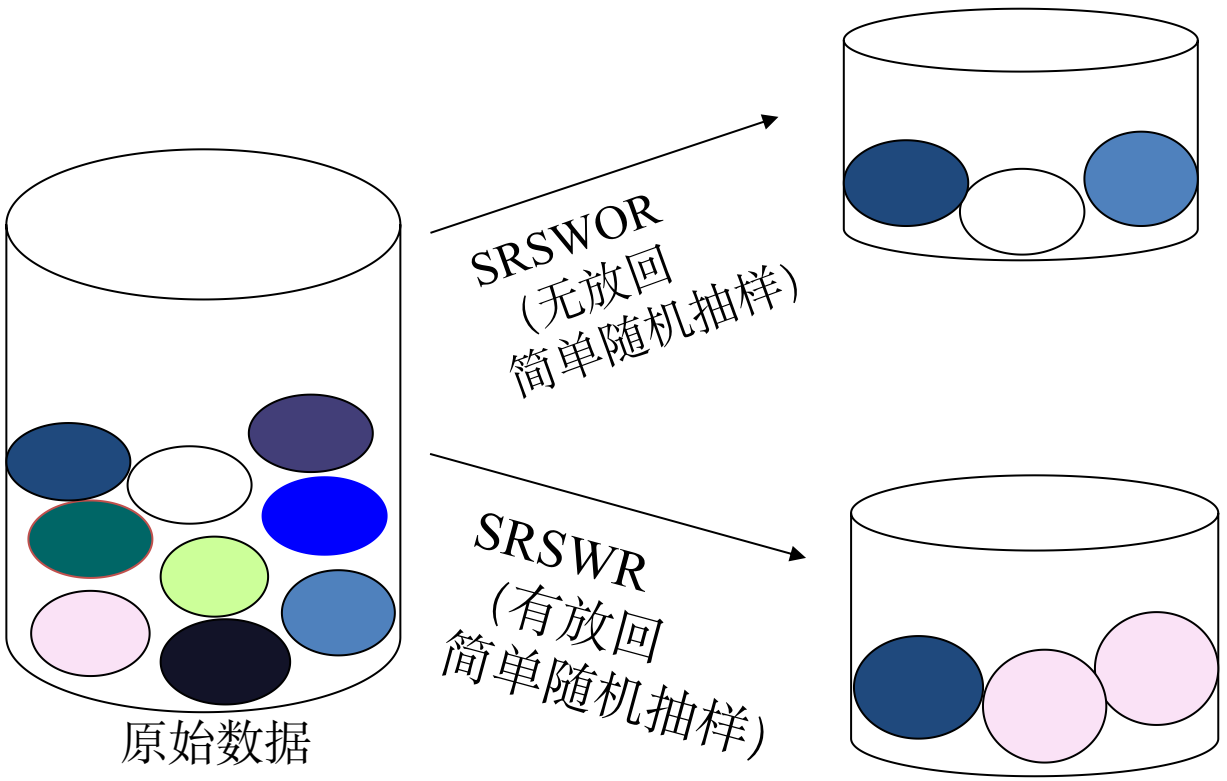
- 作用
  - 允许用数据的较小随机样本子集，表示较大的数据集；
  - 压缩数据量，允许在计算能力有限的情况下使用复杂度高的挖掘算法；
  - 解决样本不均衡的问题，比如异常检测中对与少数类采用“过采样”；

# 抽样方法


典型抽样方法
随机抽样
简单随机抽样
无放回
有放回
系统抽样（等距抽样
整体抽样
分层抽样
不平衡抽样
过采样
欠采样
相结合

# 简单随机抽样

- 简单随机抽样：从总体中抽取一定的样本，每个对象被抽取的概率相等；
- 两种变形：无放回/有放回

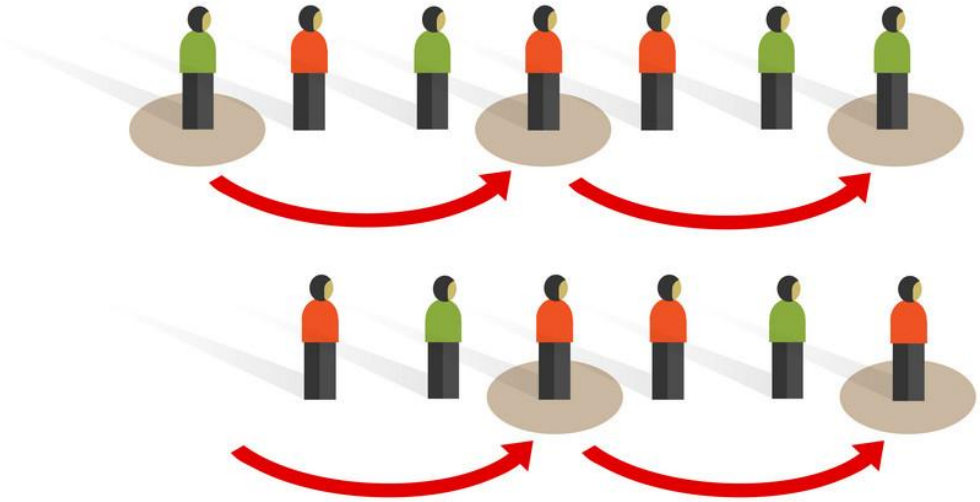



1.选中的样本立即从总体删除；
2.每轮样本被选中的概率是变化的；

1.选中的样本不从总体删除，同一样本可能被采样多次；
2.每轮样本被选中的概率是一致的；

# 系统抽样

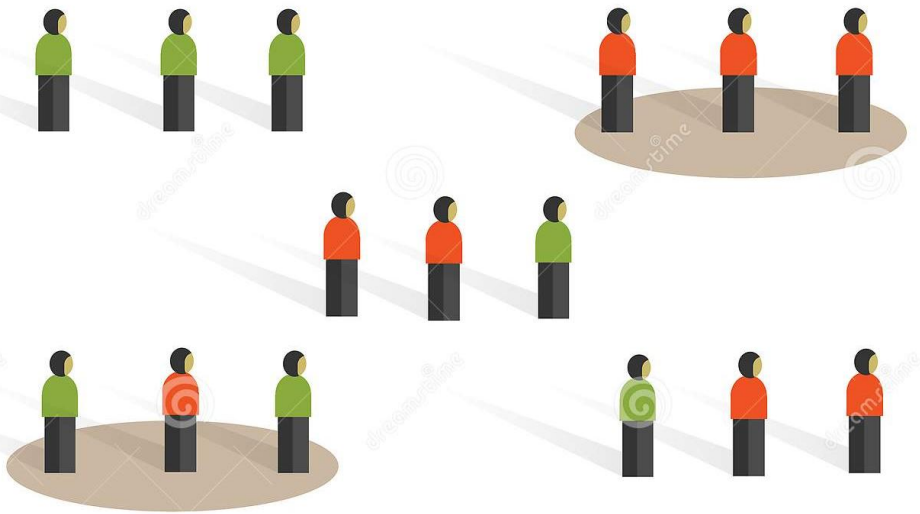
- 将总体个体按一定顺序随机编号，按编号固定间隔取样。
  - 编号顺序一定要随机打乱；
  - 间隔大小根据样本容量/总体大小而定；



- 适用性
  - 当总体规模较大或样本容量较多时，采用简单随机采样效率低，更适合系统（等距）采样。

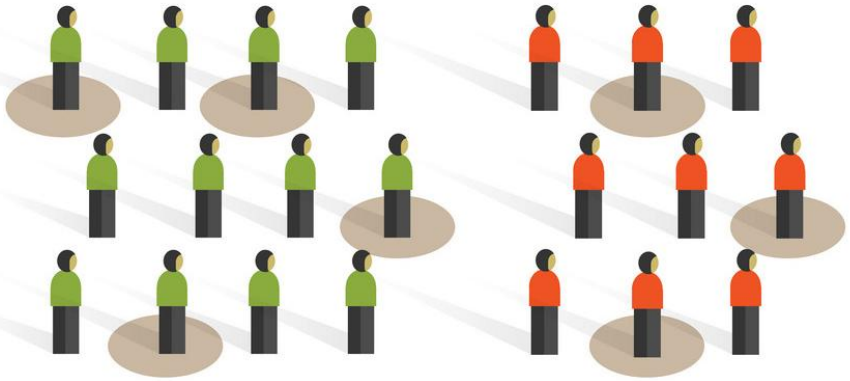
# 整体抽样

- 当总体中归并成若干个互不相交、互不重复的群，然后以群为单位抽样，随机抽取若干个群，组成抽样的总体。
  - 整体抽样要求各群有较好的代表性，即群内各单位的差异要大，群间差异要小；
  - 比如移动模式挖掘时，以“地区”为单位抽样，将北京作为中国大城市的代表；



# 分层抽样

- 当总体由几种特定类型（“层”）的对象组成时，对每个“层”分别进行随机抽样；

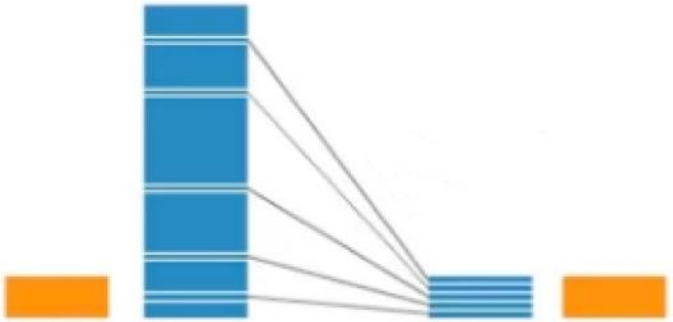


- 适用性
  - 当每种类型的对象数量差别很大时，简单随机抽样不能充分代表不太频繁出现的对象类型，适合采用分层抽样；
- 采样规模
  - 尽管每组大小不同，但从每组抽取的样本个数相同；
  - 根据每组占总体的比例确定分层采样规模；

# 不平衡抽样

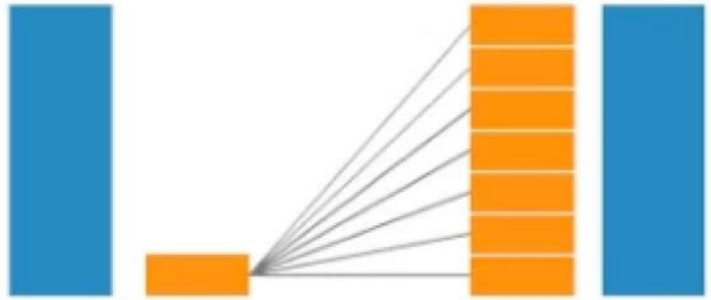
- 适用性
  - 针对数据挖掘中的不平衡数据集（如分类类别不平衡）；
- 过采样/欠采样

欠采样多数类别



原始数据　　样本

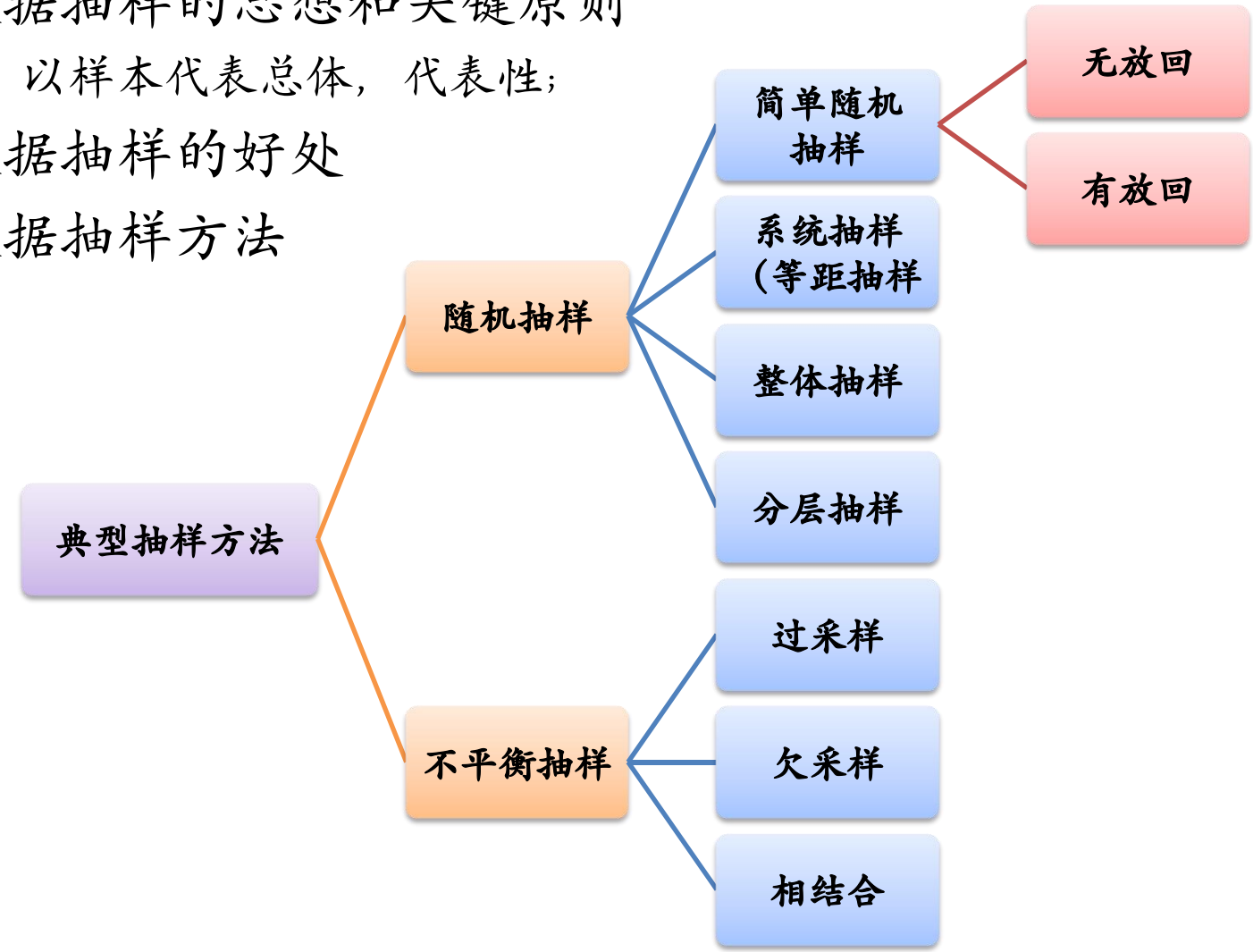
过采样少数类别



原始数据　　样本

# 小结

- 数据抽样的思想和关键原则
  - 以样本代表总体，代表性；
- 数据抽样的好处
- 数据抽样方法



# 讲授提纲

01 数据预处理概述

02 数据清理

03 数据抽样

04 **数据规约**

05 数据变换

06 数据预处理案例

# 数据规约

- 目的：降低数据维度的同时，保证基本一致的挖掘效果；

- 维度诅咒（Curse of dimensionality）：当维度增加时，数据在它对应的空间中变得愈加稀疏。
  - 数据点间的距离和密度对于聚类至关重要；
  - 子空间的可能组合会呈指数级增长，影响诸如关联分析等任务。

- 两种典型方案
  - 维规约（Dimension Reduction）：通过创建新属性，将旧属性合并表示来降低数据集的维度；
  - 属性子集选择（feature subset selection）：不创建新属性，而是选择原属性的子集；

# 维规约

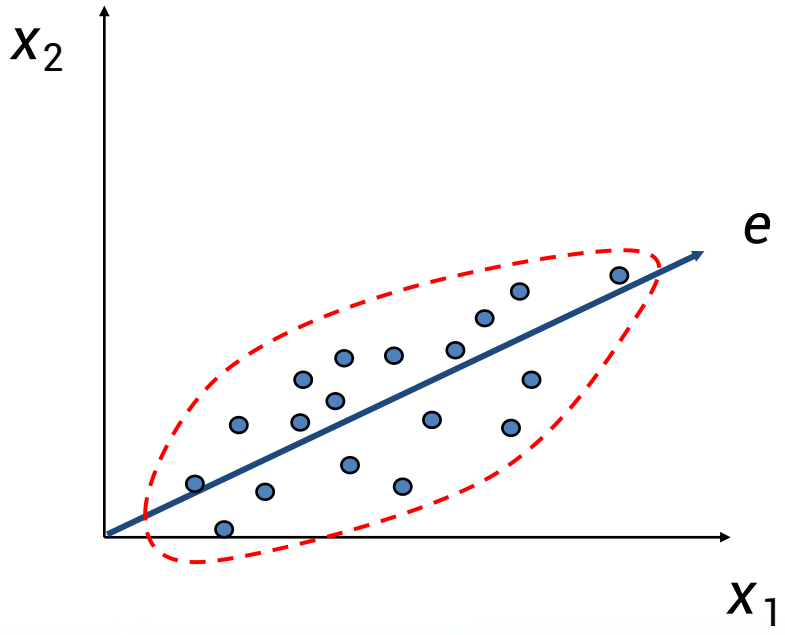
- 维规约：常采用线性代数技术，将高维空间（原属性空间）投影到低维空间（新属性空间）；

- 维规约的好处
  - 避免维度诅咒；
  - 帮助去除不相关特征和减弱噪声；
  - 减少数据挖掘的时间和空间需求；
  - 使得数据可视化更轻松；

- 典型技术：主成分分析（Principle Component Analysis, PCA）

# 主成分分析

- 主成分分析：通过正交变换，将多个可能存在相关性的变量（原属性）转换为少量线性不相关的变量（新属性），以达到降维度的目的。
  - 转换后的新属性叫主成分，是原属性的线性组合；
  - 新属性两两之间彼此正交；
  - 目标：尽可能捕捉原数据的最大方差；
  - 仅适用于数值属性。

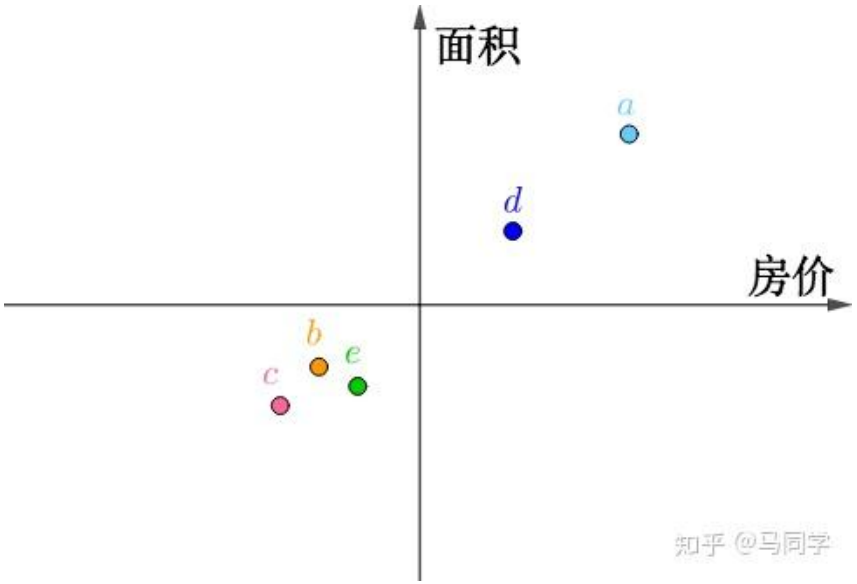


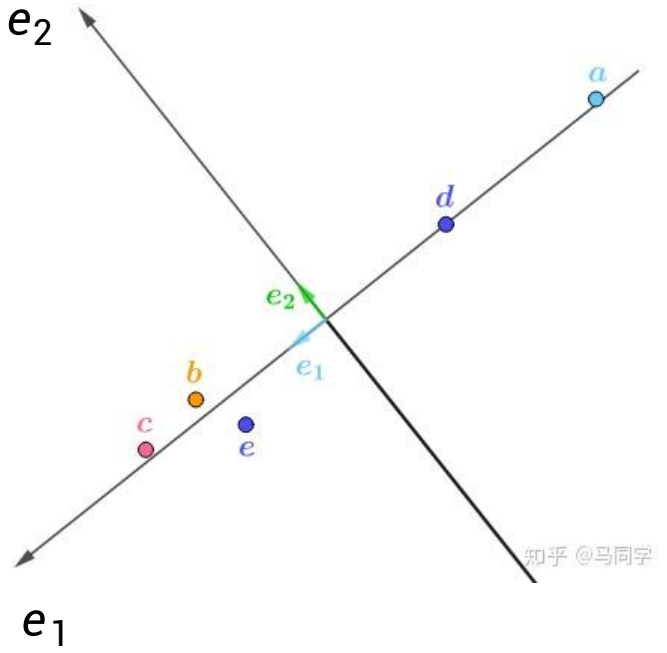

找出一个能保留原属性最大方差的投影方向

# 主成分分析：以房价为例

| | 房价(百万元) | 面积(百平米) |
|---|---|---|
| $a$ | 5.4 | 4.4 |
| $b$ | −2.6 | −1.6 |
| $c$ | −3.6 | −2.6 |
| $d$ | 2.4 | 1.9 |
| $e$ | −1.6 | −2.1 |

| | 主元1 | 主元2 |
|---|---|---|
| $a$ | −6.94 | 0.084 |
| $b$ | 3.02 | 0.364 |
| $c$ | 4.42 | 0.204 |
| $d$ | −3.05 | −0.006 |
| $e$ | 2.55 | −0.646 |



# 属性（特征）子集选择

- 降低数据维度的又一种方法，用于去除以下属性：
  - 噪音属性：属性取值基本一致，方差很小；
  - 冗余属性：与其他属性存在高度相关，没有新的信息；
  - 无关属性：与学习任务不相关，如学生ID对于GPA预测。

# 特征选择方法

- 过滤法（Filter）：用一系列统计指标筛选出部分特征，特征选择与机器学习算法无关；
- 包裹法（Wrapper）：选出多个可能的特征子集，基于机器学习算法指标评估特征子集的质量（特征子集包裹机器学习）；
- 嵌入法（Embedding）：特征选择与机器学习模型融为一体，基于统一的优化目标同时学习模型和筛选特征；

# 小结

- 数据规约：降低数据维度的同时，保证基本一致的挖掘效果；
- 维规约（Dimension Reduction）
  - 通过创建新属性，将旧属性合并表示来降低数据集的维度；
  - 主成分分析：基于正交变换寻找捕捉原数据最大方差的主成分；
- 属性子集选择（feature subset selection）：
  - 不创建新属性，而是选择原属性的子集；
  - 去除噪音、冗余、无关属性；
  - 过滤法、包裹法、嵌入法。

# 讲授提纲

01 数据预处理概述

02 数据清理

03 数据抽样

04 数据规约

05 **数据变换**

06 数据预处理案例

# 数据变换

- 通过特定函数将原属性取值映射到新的函数空间。


- 典型方式
  - 离散化
  - 规范化（或标准化）
  - 简单函数变换

# 离散化

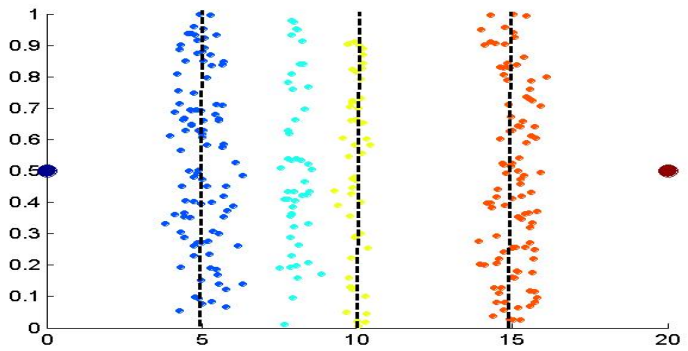
- 将连续数值属性映射到离散区间；
- 离散化的原因
  - 一些算法只接受离散化的输入；
  - 降低数据规模。
- 离散化技术
  - 等宽离散化（Equal-width discretization）
  - 等深离散化（Equal-depth discretization）
  - 基于聚类的离散化（Clustering-based discretization）

# 等宽离散化

- 将原属性取值范围（min，max）分成$N$个宽度相等的区间，每个区间的宽度 $L$ 记为：

$$L = \frac{max - min}{N}$$

- 特点
  - 最直接有效的离散化方式，但不适用于有偏的数据；
  - 对于$N$和异常点敏感；

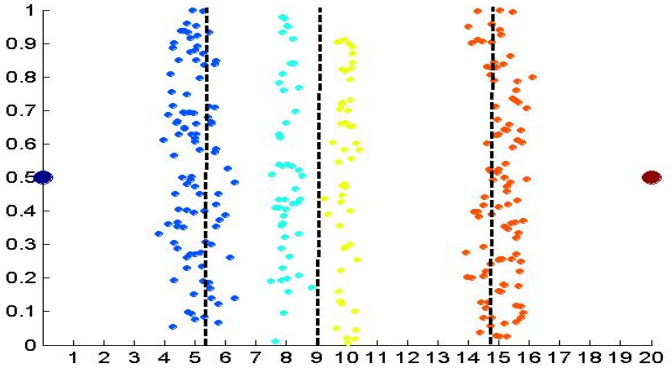



等宽离散化

# 等深离散化

- 将原属性取值范围分成$N$个区间，使得每个区间包含个数相等的数据点；
- 又叫等频（Equal-frequency）离散化。

- 特点
  - 具备较好的数据伸缩特性，更能应对有偏的数据；
  - 对异常点不那么敏感；



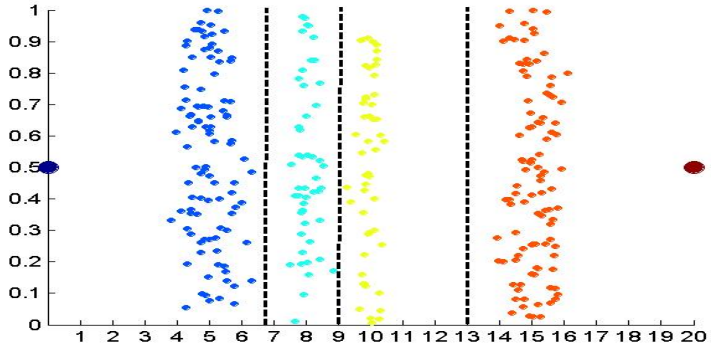

等深离散化

# 基于聚类的离散化

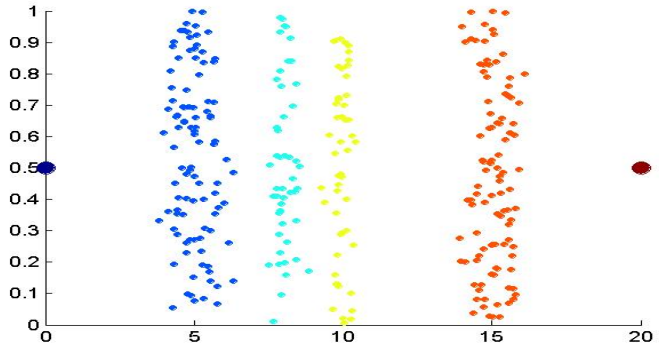
- 采用聚类技术将原属性取值范围划分为几个聚簇（区间）；

- 特点
  - 能考虑数据点的分布和邻近性，同一区间数据点尽可能临近，不同区间数据点尽可能分离；
  - 能产生高质量、符合人类感知的离散化结果。

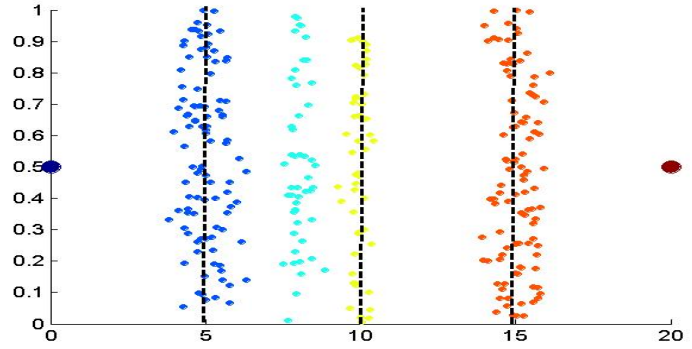
- 典型技术：基于K-Means的离散化



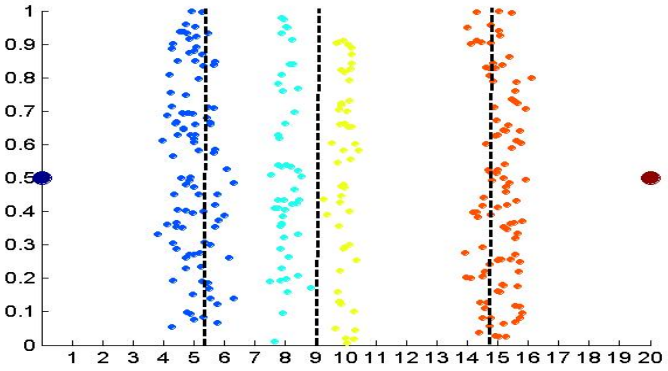
# 离散化方法图示



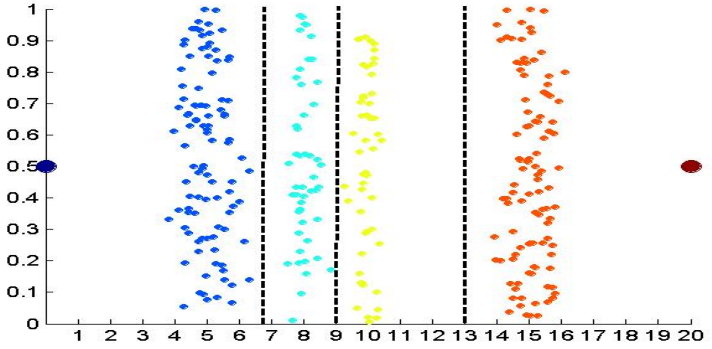
原属性取值



等宽离散化



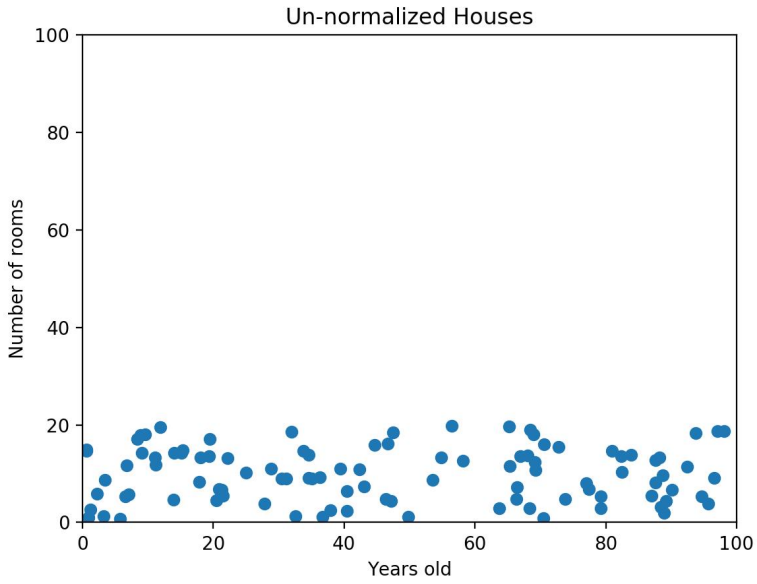
等深离散化



基于K-means的离散化

# 规范化

- 将不同属性的取值放缩到同一个较小的、指定的区间；
- 规范化的原因
  - 不同属性使用不同的量纲；
  - 取值范围大的属性占主导作用、取值范围小的属性变得不重要。

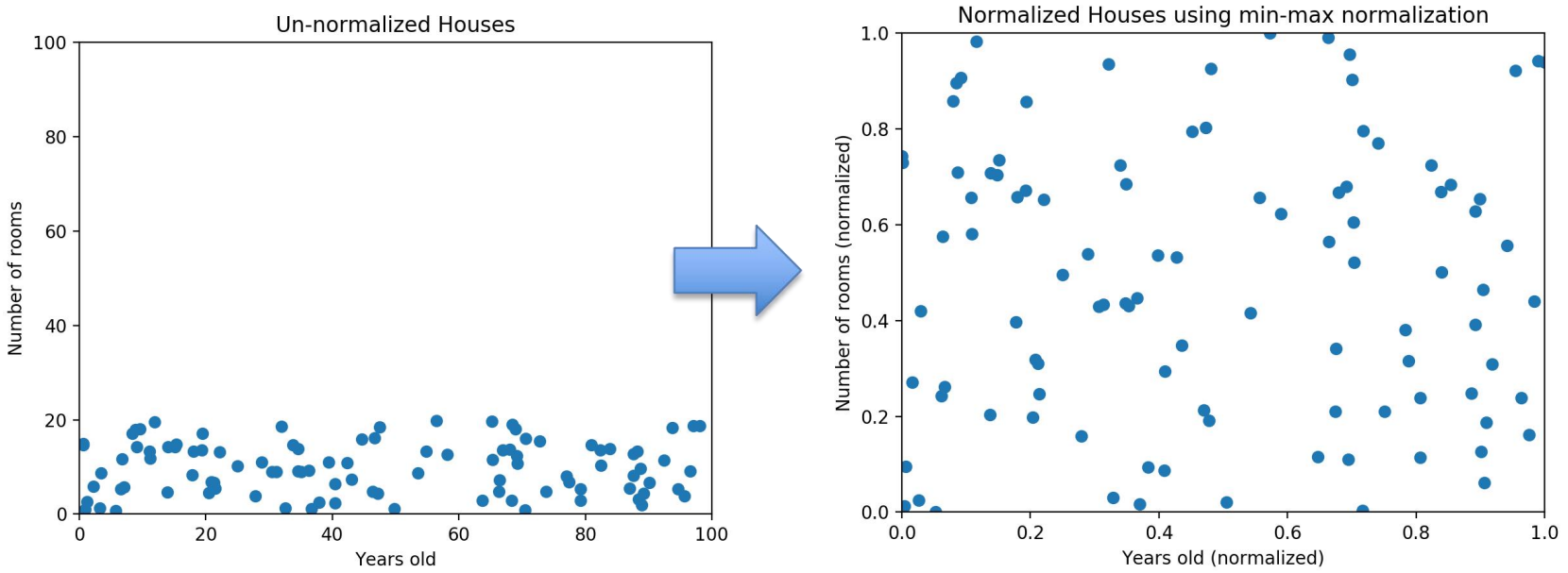



- 规范化技术
  - 最小-最大标准化（min-max normalization）
  - 正态规范化（z-score normalization）

# 最小-最大标准化

- 假定某属性原来的取值范围是$(min, max)$,变换后取值范围是$(min', max')$,则原属性取值$v$变换为:

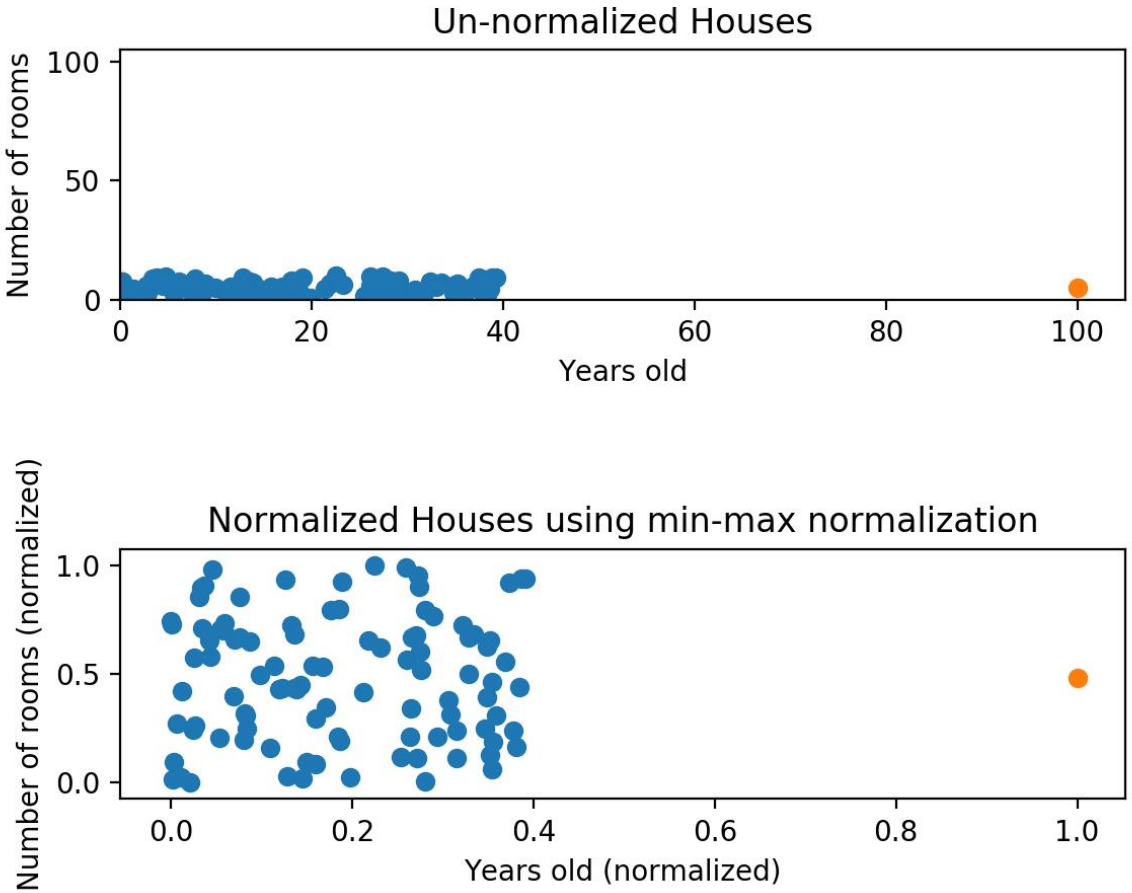
$$v' = \frac{v - min}{max - min}(max' - min') + min'$$




https://www.codecademy.com/articles/normalization

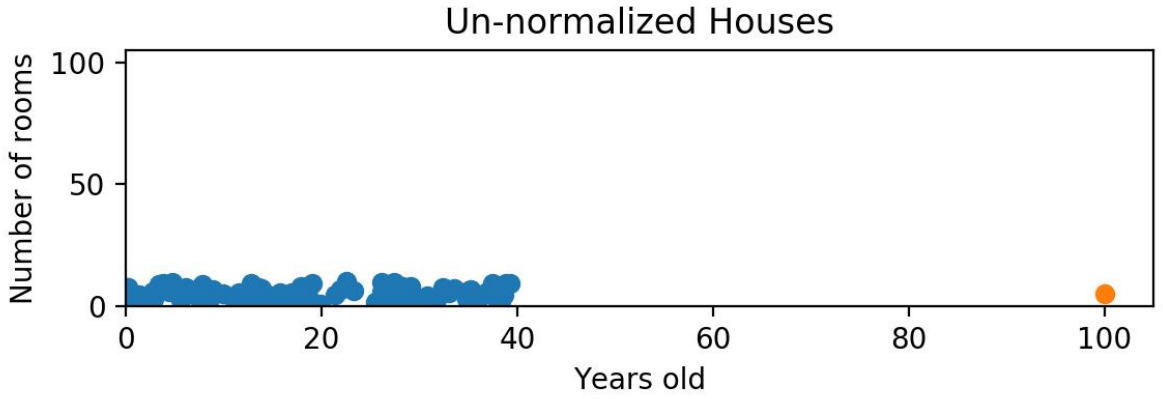
# 最小-最大标准化

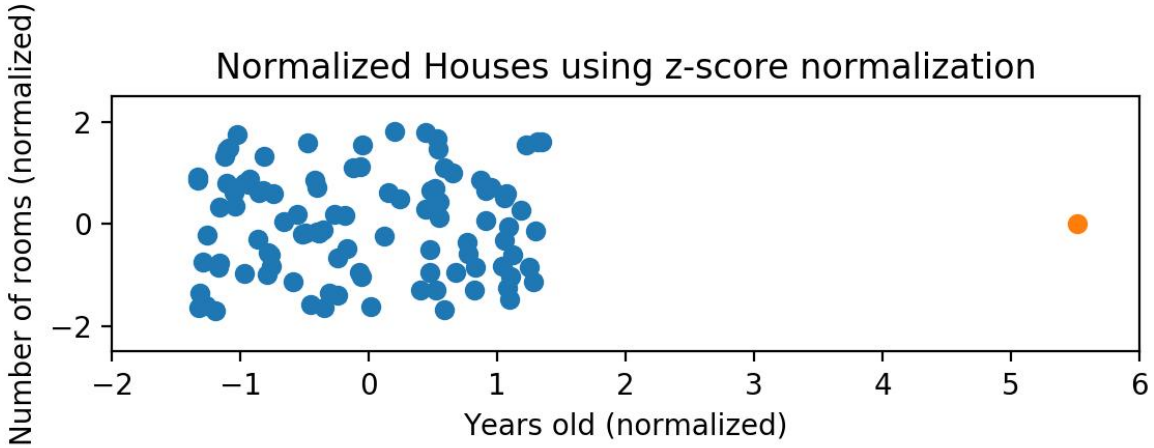
- 思考：考虑异常点的情况下，最小-最大标准化表现如何?



# 正态规范化

- 基于属性$A$的均值$\bar{A}$和标准差$\sigma_A$进行规范化：$v' = \dfrac{v - \bar{A}}{\sigma_A}$



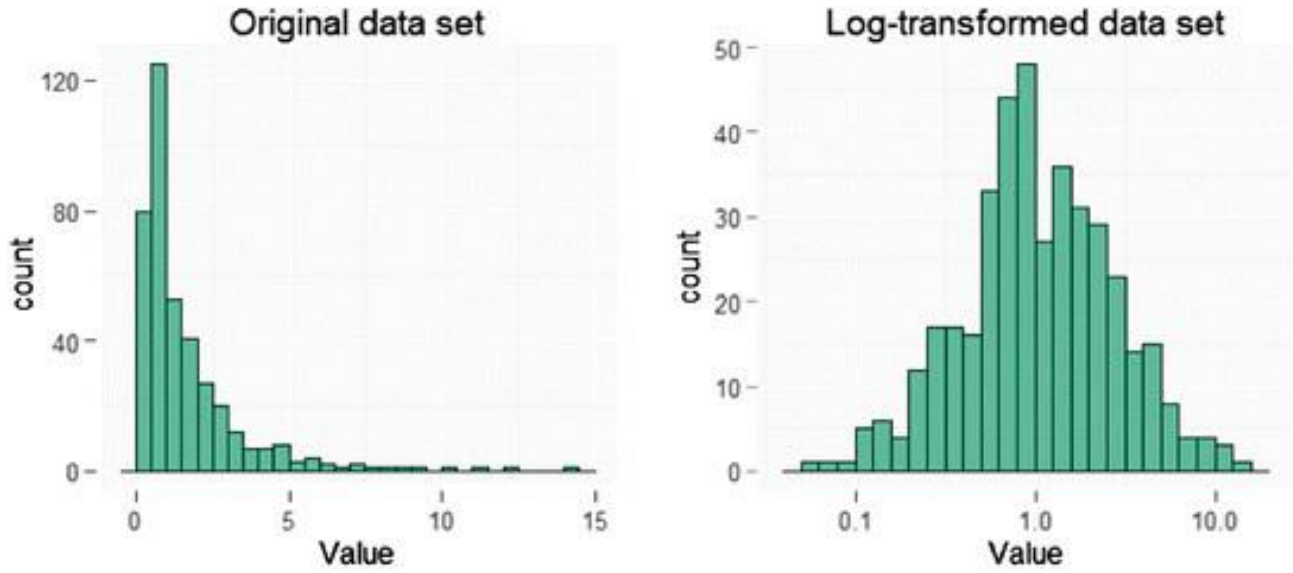



# 规范化方法对比

| 最小-最大规范化 | 正态规范化 |
| --- | --- |
| 可限定取值范围$(min', max')$ | 无法限定取值范围: $(-\infty, +\infty)$ |
| 均值、标准差不定 | 变换为标准正态分布，均值为0，标准差为1 |
| 对异常点敏感 | 能较好的应对异常点 |
| 作用：映射到特定范围，如图像处理中RGB颜色 (0，255) | 作用：无量纲化，适用于距离计算和利用梯度下降寻优的模型（如逻辑回归、SVM) |

# 简单函数变换

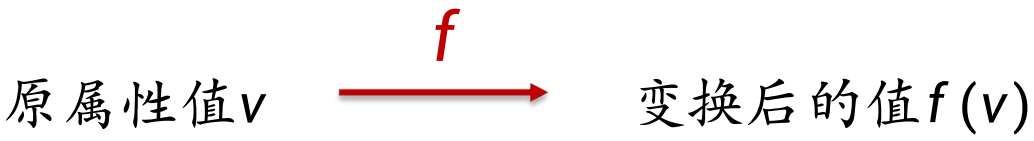
- 利用简单的数学函数，实施变量变换；
  - 如$\log x, \sqrt{x}, 1/x$,常将不具有正态分布的数据变换得具有正态分布；




  - 注意：谨慎使用变量变换，因为改变了数据特性；

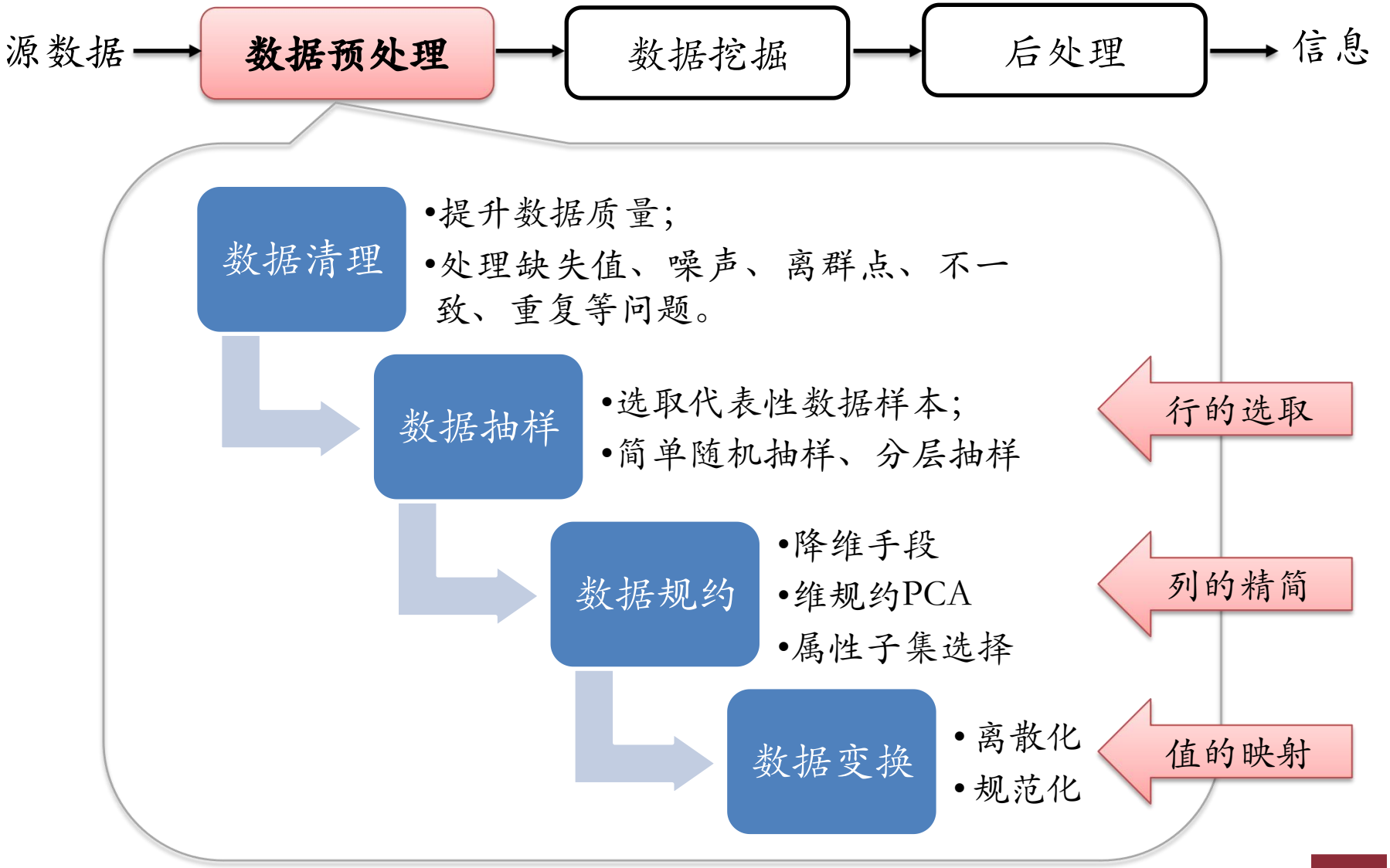
# 小结

- 数据变换：通过特定函数将原属性映射到新的函数空间。




- 离散化：等宽、等深、聚类离散化
- 规范化：最小-最大、正态规范化
- 简单函数变换：使用$log x$、$\sqrt{x}$,等， 改变数据分布特性

# 本章梳理总结


源数据
数据预处理
数据挖掘
后处理
信息
数据清理
•提升数据质量；
•处理缺失值、噪声、离群点、不一致、重复等问题。
数据抽样
•选取代表性数据样本；
•简单随机抽样、分层抽样
行的选取
数据规约
•降维手段
•维规约PCA
•属性子集选择
列的精简
数据变换
• 离散化
• 规范化
值的映射